Iowa Open Data Handbook

June 1, 2015

data.iowa.gov

EXPLORE • DISCOVER • SHARE

1 Contents

2 Introduction

When the Iowa Legislature passed the Taxpayer Transparency and Taxation Disclosure Acts (Iowa Code Chapter 8G) in 2011, their goal was to ensure Iowa taxpayers have easy access to:

- State and local government tax rate information,
- Details on state spending, and
- Data showing results achieved with taxpayers' dollars.

These goals build on Accountable Government Act (AGA) (Iowa Code Chapter 8E), passed more than a decade earlier. The AGA required state agencies to provide the "widest possible" dissemination of performance measures and performance targets based on data used by the agency to evaluate its performance. The AGA's intent was to enable state agencies to effectively and efficiently respond to the needs of Iowans and continuously improve state government performance, and for citizens to have a better understanding of how agencies perform.

At their core, Taxpayer Transparency Act, Taxation Disclosure Act and AGA intend to help citizens answer questions like:

- Which businesses in my community received permits or licenses last year?
- Are we paying more property taxes than neighboring communities?
- Have crime rates increased or decreased in my community?
- How many crashes occurred on Iowa roads last year?
- How does graduation rates in my school district compare to others?

It is now possible to provide citizens with a self-service tool to find the answer to questions such as these by releasing the data compiled by state agencies. The key benefits of making data more readily accessible include:

- Improving the public's understanding of the cost and purpose of government services
- Improving governmental accountability and public participation
- Leveraging data held by different agencies by connecting datasets and finding new insights
- Eliminating redundancies by allowing the access of data in one place
- Improving decision making by better informing people with data
- Creating more efficient and proactive process for open records requests
- Encouraging innovative ideas (e.g., web applications) that enhance the lives of our citizens

- Increasing economic activity by generating new and rich content through new applications and services

3 Handbook's Purpose

This handbook is intended to serve as a guide for state agencies contributing to the data.iowa.gov. It outlines who is required to use the portal, and provides guidelines for identifying, reviewing, prioritizing and preparing state data for publication, and maintaining that data once published. It also highlights methods for communicating data, and points to resources to more effectively and efficiently use data.iowa.gov.

For the purposes of this handbook, the term "agency," shall refer to any state department, office, board, commission, bureau, division, institution, or public institution of higher education.

4 Open Data Portal - data.iowa.gov

The Iowa Department of Management and the Office of the Chief Information Officer implemented **data.iowa.gov** and companion websites to meet the Legislative requirements of the Taxpayer Transparency and Taxation Disclosure Acts, as well as, leverage new technology to streamline state agencies' abilities to comply with the AGA. However, beyond the basic requirements, the websites foster a fundamental shift in how agencies and institutions share information by:

- Advancing interoperability so data can be more easily shared both inside and outside of state government
- Enhancing the discovery of data by citizens, businesses, state agencies, and others
- Ensuring non-technical state agency staff can quickly and easily publish and update data without IT support
- Providing simple tools to turn raw data into a meaningful interactive experience
- Making it easier for citizens to interact with government and find answers to their questions

5 Participation

The Taxpayer Transparency Act, specifically Iowa Code Section 8G.3, makes participation **all inclusive** by defining agencies as a state department, office, board, commission, bureau, division, institution, or public institution of higher education – including elective offices of the Executive Branch, agencies of the General Assembly, and the Judicial Branch. Programs and activities that are administered by or involve more than one agency are also included. Beyond budget and expenditure data provided by the Department of Management and Department of Administrative Services, your agency is required to provide information on

performance outcomes achieved by budgeted activities. You are also encouraged to provide other information you deem important for a budgeted activity.

6 Data Governance

Data governance outlines a set of processes to ensure that our data assets published on data.iowa.gov are formally managed so that data.iowa.gov and companion websites/applications (e.g. checkbook.iowa.gov) are resources the public can depend upon. Key roles in the governance structure include the state data administrator, agency data coordinators, agency data stewards, and technical support staff. Although this is a centrally managed portal, ownership, management, and responsibility for the individual datasets your agency publishes remains with your agency.

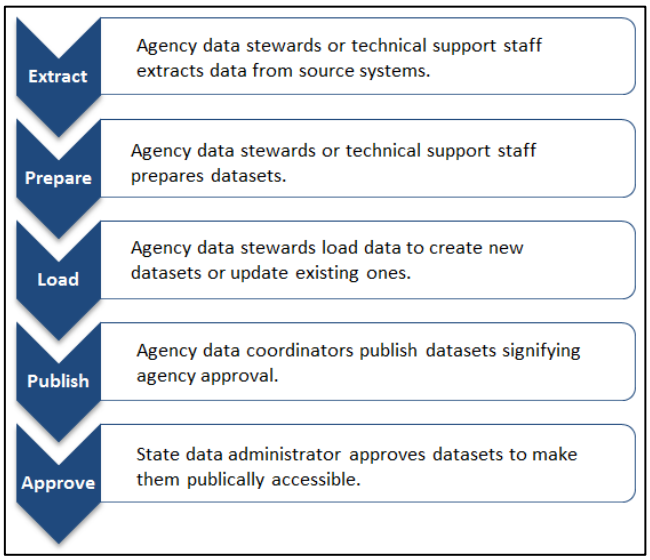


Figure 1 – Basic Dataset Approval Workflow

6.1 State Data Administrator

The Director of the Department of Management shall designate an administrator to serve as manager of the state's transparency websites, and state-wide open data initiatives. The administrator's responsibilities include:

- Overseeing the operation and evolution of data.iowa.gov and companion websites in cooperation with the Office of the Chief Information Officer
- Overseeing the development and updates of Agency Data Plans, and developing a state-wide data dashboard
- Providing support for Publishing Datasets, and Creating Data Views
- Developing best practices for agencies in sharing and communicating open data
- Evaluating and implementing practices to proactively use data collected for state-wide performance management and strategic decision making
- Managing agency accounts and roles
- Managing and updating homepage content
- Moderating comments made and views created on data.iowa.gov

6.2 Agency Data Coordinators

Each agency director shall designate an individual to coordinate your agency's efforts related to data.iowa.gov. The data coordinator should have knowledge of resources and data used by your agency. The agency data coordinator's responsibilities include:

- Completing and subsequently updating your Agency Data Plan and obtaining your agency's senior management approval
- Working with data stewards to ensure data meets your quality assurance requirements and contains relevant information
- Reviewing and publishing new and updated agency datasets on data.iowa.gov
- Addressing feedback received on your agency's published data
- Serving as liaison between your agency and the state data administrator
- Conveying your agency's specific needs of the Iowa transparency websites

6.3 Agency Data Stewards

Agencies will likely have multiple data stewards within your agency's programs that can assist your agency data coordinator. These individuals are typically the ones who collect, manage, maintain and/or analyze your agency's data. The data steward's responsibilities include:

- Defining the relevant data to include in your agency's datasets
- Creating/updating requirements to maintain data quality and integrity
- Ensuring data documentation is developed and maintained (e.g. metadata, data dictionary)
- Importing New Datasets and Maintaining Datasets
- Creating Data Views of datasets to better communicate information
- Addressing specific feedback related to your agency's data

6.4 Technical Support Staff

Agencies are likely to have one or more individuals who provide technical support to your data stewards. While they have variety of technical responsibilities related to your agency's data sources, their responsibilities related to data.iowa.gov include:

- Retrieving required data from the source system (e.g. database, application, file) in the most efficient manner
- Sorting, filtering, joining and aggregating data to ensure it meets requirements (see sections on Preparing Datasets, Publishing Quality Data, Protecting Confidential and Sensitive Data), and is in an appropriate File Formats for publishing on data.iowa.gov
- Cleaning the data to resolve inconsistencies and fix problems in your source data
- Creating and executing workflows to export your data from the source system in support of Automating Updates where appropriate

6.5 Data Ownership

Your agency retains ownership over the datasets that you publish. All data and datasets remain the property of the originating Agency and public users acquire no ownership rights to your agency data or datasets. The datasets published on data.iowa.gov or any Iowa transparency website becomes a public resource available to anyone with access to the Internet. The public use of the datasets may include development of applications. In this case, the developers retain all intellectual property ownership in their applications, excluding the agency data itself, whose ownership continues to reside with the agency.

6.6 Responsibility for Quality, Integrity and Security

Your agency is responsible for all aspects of the quality, integrity, and security of dataset content published. You do not relinquish control of your data when a dataset is published on data.iowa.gov or any Iowa transparency website. Your agency is responsible for ensuring that all of your published data has been reviewed by appropriate management for confidentiality, privacy, security and all other content limitation issues before the data is published.

7 System Accounts

State agency staff must have an account on data.iowa.gov to contribute to the open data portal. Each account will be granted the following privileges on data.iowa.gov:

- View site-wide analytics (Administrative Dashboard)
- Importing New Datasets and Maintaining Datasets
- Creating Data Views on datasets (maps, charts, etc.)
- Sharing Datasets and Changing Owners

7.1 New Account Requests

Once your agency data coordinator and agency data stewards have been identified you will need to submit an email request to the state data administrator to set up accounts and privileges in data.iowa.gov for the individuals. ***Do not create an account directly***. The following information should be provided for each agency staff person requiring an account:

- Name of the department, division, bureau or program the individual is representing, and recommended display name for the account (visible to the public). The display name should represent the organizational unit the individual is publishing data for. It is recommended the display name be less than 30 total characters including spaces.

Tip

Your agency may have already created a profile image for existing social media accounts that can be used for this purpose

Agency data coordinators display name should be the only account associated with the name of the state agency.

- First and last name of the individual requiring the account.
- Work email for the individual requiring the account.
- Your department, division, bureau or program logo to use as the account's profile image. The logo (or portion of it) should be cropped so it is square. You may elect to use another image that represents your agency's mission if a logo does not exist or cannot be modified to meet the profile image specifications.

7.2 Changing Account Settings

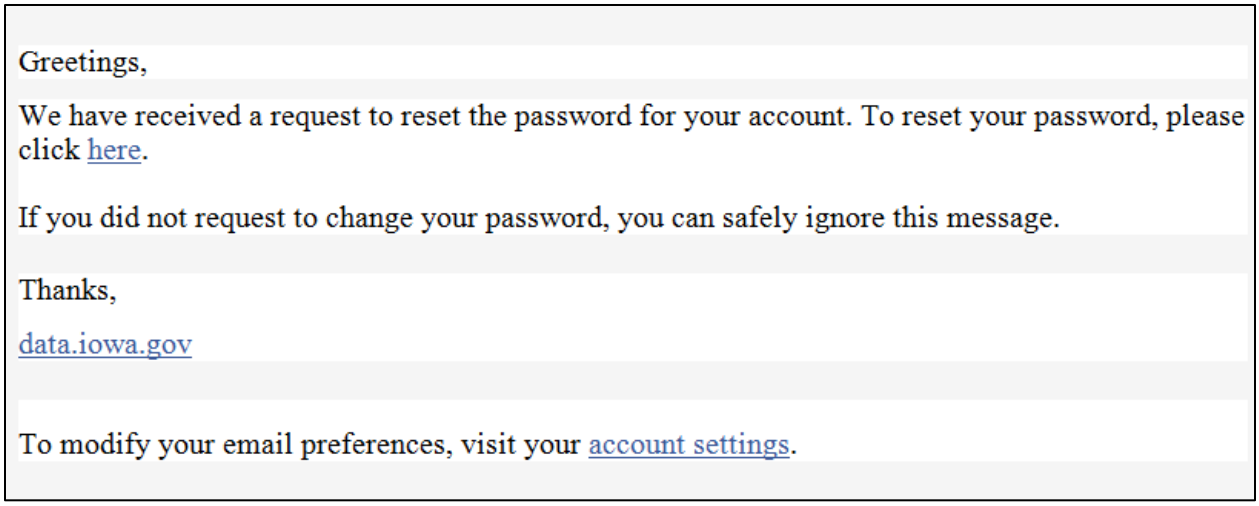
Once you sign into your account, you will be taken directly to your account's dashboard. You may access your dashboard at any time by clicking the "Hello" link that includes your organization's name in top right hand corner. Once on your account's dashboard, you should click "Edit Account Settings" under the menu bar. This will take you to a web form where you can:

- Change the email associated with the account
- Change the password associated with the account
- Subscribe to email notifications

All users should change your password from the default once a new account is created. If duties change or staff turnover, the email associated with the account should also be updated to use the email of the individual who will assume responsibility for the account.

7.3 Resetting Passwords

When you click the "Sign In" link in the top right hand corner of the page – above the search box – you will see a "Forgot Password?" link on the web form used to sign into data.iowa.gov. To reset your password, you will need to enter your email and click the red "Reset Password" button. This will send an email the account on file with link to reset the password, see figure 2 below.



Greetings,

We have received a request to reset the password for your account. To reset your password, please click here.

If you did not request to change your password, you can safely ignore this message.

Thanks,

data.iowa.gov

To modify your email preferences, visit your account settings.

Figure 2 – Email received to reset password

8 Agency Data Plan

In order for you to effectively utilize the transparency websites, and better manage publishing your agency's data, you need to understand what data you have and determine if the data are of high value and are ready to publish. Your agency plan should include an inventory of key data, impact and difficulty assessments, priorities and schedule. More information on drafting your agency data plan can be found in Appendix A. Drafting Your Agency Data Plan.

Once your agency data plan has been drafted and approved by your agency's senior management, it should be submitted to the state data administrator with the Department of Management.

Agency data plans should be updated annually. Agency data coordinators should submit plan updates to the state data administrator with the Department of Management by June 15th for the upcoming fiscal year.

9 State Data Dashboard

The state data administrator will create a dashboard based on information provided in agency data plans to promote accountability. The dashboard will be made available to the public, and will track agency progress towards publishing datasets on data.iowa.gov. The dashboard will display by Agency:

- Scheduled dataset releases
- Inventory of datasets
- Dataset releases behind schedule

10 Preparing Datasets

A dataset is a named collection of related records containing data organized or formatted in a specific or prescribed way, often in tabular form. There are a number of things you should consider when preparing datasets to publish to ensure they are relevant to potential users.

Your dataset is ready when it:

- ✓ Provides sufficient level of detail
- ✓ Contains headings
- ✓ Provides a key field where available
- ✓ Includes numeric, contextual, and categorical data
- ✓ Sums across columns and not down rows
- ✓ Includes addresses or other location data if available
- ✓ Is based on queries with proper joins
- ✓ Logically orders columns
- ✓ Contains 30 or fewer columns

10.1 Granularity

There are no hard and fast rules about what level of detail is sufficiently granular to add value to a dataset. Users of data.iowa.gov may come from a variety of fields and specialties, including developers, and academic and other government users who can envision a use for the raw data not originally anticipated. Raw or very granular data also provides more options for presenting or summarizing the data (i.e. looking at it in different ways). By providing granular datasets, you facilitate making data more meaningful to a wider cross section of the public. However, safeguards to prevent the disclosure of personally identifiable information and other confidential data may limit how granular a dataset may be. See the section on Protecting Confidential and Sensitive Data for more information.

10.2 Joins

Many datasets will be created with the use of complex queries involving joins of two or more tables. Joins have the potential to affect the number of rows returned and ultimately the accuracy of the dataset. To help avoid problems, it is a good practice to begin joins with the most granular table or query.[1] Often times, this table or query will contain the key Numeric Data for your dataset that you want citizens to be able to summarize (e.g. count of offenders in prison, sum of payments made, etc.). By beginning joins on this table or query, you can better trust that any aggregates on your primary table or query will be accurate. Oftentimes, the tables or queries you join to the primary table or query will contain Contextual Data or Categorical Data (e.g. institution, payee, etc.), which facilitates summarizing the data.

Tip

Understand type of join; Count rows before and after any join

[1] Making the most granular table/query the primary table will often facilitate creating 1 to 1 joins, where one record in the primary table/query matches to one record in the joined table/query; or Many to 1 joins, where many records in the primary table/query matches to one record in the joined table/query.

Problems can occur if one row in your primary table or query can match up with multiple rows in a joined table or query. This can lead to duplicate numeric values, which would cause inaccuracies with any aggregate function.

10.3 Column Identifiers

Your agency must include column identifiers (i.e. headings) in the first row of your file for each column containing data. When creating headings, you should:

- Keep them short and meaningful
- Keep alpha characters lower case
- Avoid the use of symbols in headings (e.g. & or %)
- Use underscores in place of spaces in headings
- Ensure each column heading is unique

This will allow the system to use the headings you provide as the API field name for the column, and will facilitate Automating Updates in the future. In the system, you will have an option for Changing Column Properties and provide more descriptive or human readable labels and definitions – both of which can further expand users' understanding of your data.

10.4 Row Identifiers

When possible, you should include a column that uniquely identifies each row in the dataset. This value should be unique and not be contained in more than one row. Values are typically numbers or some alphanumeric code. Providing row identifiers can help facilitate programmatically and automatically updating datasets. A row identifier will also allow developers to use this column to power applications. That way if columns are deleted or added, developers are ensured that the applications built off of the row identifier will not break. It can also help you later on where specific records in the dataset need to be corrected or replaced due to errors.

10.5 Numeric Data

It is highly recommended that you use numeric and date data types whenever possible. This allows the system to perform calculations on numeric data to create charts, and use dates to build calendars that is not possible with plain text. If you have rows (or records) with missing values, due to either data collection issues, or for confidentiality reasons, it is recommended that you create another column to flag or tag the record and briefly define the reason for missing data.

10.6 Contextual Data

It is important that you include information in your dataset that give numeric data context. For instance, a dataset containing payment transactions is far more meaningful if the payee, transaction date and expense description is also provided.

10.7 Categorical Data

Categorical data is often broader groups or categories for the contextual data previously noted. Grouping data is often necessary to effectively calculate (e.g. sum or average) numeric data and provide meaningful and understandable summaries for citizens. While some users will want access to the "raw" data, most citizens will find visual summaries, such as a chart or map, far more helpful. Categorical data facilitates Grouping Data. Reference tables in source systems often provide the categorical data you will want to include. You can also review your data to see if column headings can be transformed into data values (e.g. store dates in rows, not columns).

10.8 Summary Data

It is best to only store the raw data and leave calculations up to the system. So, don't include those rows providing subtotals or totals for a group of records (i.e. rows). This makes your data more flexible by allowing it to be summarized in many different ways. However, having a column that sums or totals values from other columns is fine.

10.9 Location Data

Being able to associate the records (i.e. rows) in your dataset to a geographic location can allow you and users of data.iowa.gov to build maps using your data. Records can be tied to a geographic location using a full or partial addresses, geographic codes, or more complex geometric structures.

Addresses are likely the most common form of location data collected by agencies. Including address information, where possible, such as a U.S. street address (excluding PO Box), city, state and zip code, will allow the system to geocode the address and display it on a map, as shown in Figure 3. It is recommended that each part of the address be in a separate column.

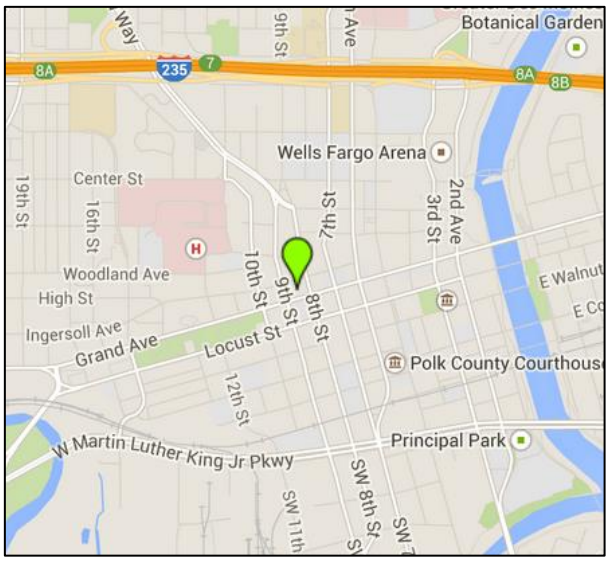


Figure 3 – Location Data. A geocoded address can be displayed as a point on a map.

Zip code, city and state, city and zip code, and state can also be geocoded on their own if a full address is not provided. If your data

has already been geocoded, consider including separate columns for longitude and latitude (in decimal degrees), so the geocoding process can be skipped.

It is also recommended where appropriate to include standard geographic codes for counties, county subdivisions cities, etc. (e.g. Federal Information Processing Series (FIPS) Codes or Geographic Names Information System (GNIS) Identifiers).

More complex geometric structures such as lines and polygons, can also be handled, see the section on File Formats.

10.10 Column Order

The following provide general rules of thumb related to how columns should be ordered within your file to provide some uniformity to how data is presented.

- Column containing Row Identifiers should be the left most column in the file
- Columns containing Categorical Data and Contextual Data should be placed to the left of columns containing Numeric Data
- Columns containing alphanumeric codes should be placed to the immediate left of columns containing the labels for the codes (e.g. column containing department numbers should be placed to the left of the column containing the department names)
- Columns containing Categorical Data that are hierarchical should be adjacent to one another. Additionally, the parent column should be placed to the left of the child column (e.g. column containing department names should be placed to the left of column containing division names)
- Aggregate columns totaling Numeric Data from other columns should be placed to the right of columns included in aggregate total

Columns may be reordered after import. However, creating files with columns ordered in the manner above will simplify updating your data in the future.

10.11 30 Columns or Less

You should consider restructuring your dataset if it has over 30 columns. Reducing the number of columns may be achieved by stacking your data, or splitting the data into two or more datasets.

11 Publishing Quality Data

Your agency should implement a quality assurance process (i.e. document data checking and review steps) for datasets you intend to publish on data.iowa.gov. How rigorous your quality assurance process is will be highly dependent upon the nature of the dataset and

how it is used. This section is not intended to prescribe what quality assurance[2] should look like for your data, but rather highlight a few items to consider before publishing a dataset.

11.1 Does your dataset contain duplicate records?

Many datasets will be based on complex queries involving joins which have the potential to affect the number of rows returned if done incorrectly. For datasets that are not grouped, you should compare the number of records (or rows) in your dataset to the primary table (i.e. the most granular table with the key numeric data you want to include) before publishing your dataset. For datasets that are grouped and contain aggregated numeric values, you will want to compare values in the dataset against reports in the source system. If there are discrepancies, you should review the previous paragraph on Joins.

11.2 Does your dataset contain inconsistent data values?

If your dataset has categorical data that was not the result of a "many to 1" or "1 to 1" join[3], or the database, spreadsheet, or file where the data resides does not enforce input standards, there is the possibility that your dataset will contain inconsistent values: For example:

- "Mt. Pleasant" and "Mount Pleasant" refer to the same town in Iowa
- "Dept. of Public Health", "DPH" and "Iowa Department of Public Health" all represent the same state agency
- "M" and "Male," or "F" and "Female" refer to the same gender
- Categorical data value with leading and/or trailing spaces, and same value without
- Different addresses for the same entity

Online Resource

Create a Pivot Table to Analyze Data

If inconsistent data values are present in your dataset, aggregated numeric data grouped by inconsistent categorical data will be inaccurate. Where you think this may be the case, you should consider running a pivot table on your data or use some other method to see a list of unique values in the categorical data and identify values that represent the same category. If you have multiple values representing the same category, you will want to decide which value you intend to use, and which values you intend to replace. The effort to correct 100% of inconsistencies found may not be worth it. Any known inconsistencies that are not corrected should be highlighted in the dataset's Disclaimers - under limitations.

[2] Quality assurance encompasses everything from how your data is collected, stored, managed and assessed.
[3] With 1 to 1 joins, one record in the primary table matches to one record in the joined table. With many to 1 joins, many records in the primary table matches to one record in the joined table.

There may also be instances where the definition of categorical values has changed overtime (e.g. school consolidations, business mergers, etc.). This is likely hard to detect in the data itself, and will make it challenging to make comparisons over time using those categories. Where this is known, you should attempt to document or highlight it, as it may serve as a constraint for how the data can be used.

11.3 Is your dataset missing data?

Missing values, or attributes (i.e. contextual and categorical data) can result in misleading results. To identify areas with missing values, you should scan the dataset and note gaps in the records. You can also group and aggregate data at various levels of granularity and compare against other reports to find discrepancies. Discrepancies might indicate possible missing values. If you have missing values, you will want to determine if the missing records are localized in any way or if they are random. From there, you will want to determine what you intend to do:

- Can the missing values be ignored? Do you intend to omit records with missing data?
- Do you intend to replace the missing values with substitute values?

Any known missing values that are ignored or omitted should be highlighted in the dataset's Disclaimers - under completeness. Where substitute values are used in place of missing values, it is important to make note of it under limitations.

11.4 Does your dataset contain ambiguous data values?

Data ambiguity arises when what the data represents is not precisely defined due to incomplete or conflicting definitions, subjective differences in evaluating data, or where a record could be classified in more than one way. This can lead to data values being misinterpreted. Most issues pertaining to ambiguity will need to be addressed at the point of data collection or data entry. However, some steps can be taken to address some ambiguity. For instance:

- Abbreviations and acronyms can cause ambiguity as they can mean different things (e.g. DHS can represent two different government agencies: Department of Homeland Security at the federal level or Department of Human Services at the state level). Can you take steps to expand abbreviations and spell out acronyms before you publish your data?
- Dates with the year represented as two digits are another form of ambiguity that you should address before publishing your data. Can you convert two digit years to four digit years before publishing your data?

- Many words can have multiple meanings. Can you provide data dictionaries and/or highlight the meaning of the extensively used term in the data's metadata (see Attachments) to ensure you provide appropriate context?

11.5 Does the dataset contain the expected variables or columns?

To insure data is useful to citizens and the public, you will want to verify that your dataset contains the variables (i.e. columns) you expected it to. More details on types of data to include are covered in Preparing Datasets.

11.6 Are data values formatted appropriately?

A variety of data types need to be formatted in specific ways to be imported successfully into data.iowa.gov. Appendix C. Data Types provides more information.

11.7 Documentation

Any checks and/or corrections that need to be made data updates should be documented

12 Protecting Confidential and Sensitive Data

Data containing personally identifiable information and other confidential/sensitive data protected by state (e.g. Iowa Code Section 22.7 or other applicable Iowa Code section) or federal law (e.g. Health Insurance Portability and Accountability Act (Privacy Rule and Security Rule), Privacy Act, Patient Safety and Quality Improvement Act (Patient Safety Rule), Social Security Number Protection Act, and Family Educational Rights and Privacy Act) must have adequate safeguards in place to mitigate potential privacy or security risks, and ensure such data is not compromised.

12.1 What is Personally Identifiable Information?

The U.S. Government Accountability Office has provided a comprehensive definition for personally identifiable information (or personal information):

> "any information about an individual maintained by an agency, including (1) any information that can be used to distinguish or trace an individual's identity, such as name, Social Security number, date and place of birth, mother's maiden name, or biometric records; and (2) any other information that is linked or linkable to an individual, such as medical, educational, financial, and employment information."[4]

Personally identifiable information could also include home street addresses and emails, photographic images, etc. Personally identifiable information will be the most common

[4] GAO Report 08-536, Privacy: Alternatives Exist for Enhancing Protection of Personally Identifiable Information, May 2008, http://www.gao.gov/new.items/d08536.pdf.

type of confidential/sensitive data you will likely encounter when working to publish your data.

12.2 What other information is confidential?

While this is by no means an exhaustive list, besides personally identifiable information, the following may also be considered confidential or sensitive:

- Information that has commercial value and upon release could give advantage to the information provider's competitors (e.g. trade secrets, sales and marketing plans, blue prints, process designs, financial data, etc.)
- Information related to electronic signature, internet protocol numbers, security of electronic transactions
- Records relating to charitable giving
- Investigative and autopsy files and reports
- Work products related to litigation
- Preliminary, draft work and research materials that are not in final form
- Location of sensitive ecological sites or archaeological resources
- Information security and emergency preparedness procedures
- Information on critical assets

12.3 Are there exceptions?

Before automatically assuming data is confidential or sensitive, review any applicable laws to see if disclosure is authorized or if exceptions exist. Even if an item is generally thought of as confidential or sensitive, there may be instances where law makes exceptions or where code specifically requires the information to be made available.

12.4 Unintentional Disclosure – Small Groups

Unintentional disclosure of personally identifiable information is one of the greater challenges to publishing open data. Even when you are mindful of and have suppressed personally identifiable information, and other confidential data, your data may contain easily observable and/or unique characteristics where alone or in combination can increase the risk of unintentional disclosure by producing a small group of individuals for which they apply.[5] Some examples include:

- Very high income
- Very low income
- Unusual occupation
- Unusual health condition

[5] Checklist on Disclosure Potential of Proposed Data Releases, Federal Committee on Statistical Methodology, July 1999, http://fcsm.sites.usa.gov/files/2014/04/checklist_799.doc

- Very high age
- Unusual racial category
- Small geographic area
- Small time period

Many statisticians consider *a group of 3 to be the minimum* needed to prevent disclosure, though larger minimums (e.g., 5 or 10) may be used to further mitigate disclosure risk if deemed appropriate for the nature of the data. However, a small group by itself does not necessarily constitute disclosure. When considering disclosure risk, you should consider whether a "reasonable person" (i.e., a hypothetical, rational, prudent, average individual) with no personal knowledge of relevant circumstances could identify the individual based on publicized events or highly unique characteristics associated with the individual.[6]

12.5 Unintentional Disclosure – Group Disclosure

Unintentional disclosure could also occur where a large percentage of individuals fall within a category and makes it fairly certain who the data applies to. For example, if a dataset shows all 11th grade males at a specific school as "Below Proficient" on a given assessment, there is no question who the information applies to.

12.6 What safeguards can prevent disclosure of confidential/sensitive data?

Oftentimes, datasets do not need to contain complete or full data records to be considered valuable. Where confidential/sensitive data exists, there are safeguards you can use. A few methods are noted below. The method or combination of methods you use will depend on the disclosure risks associated with your data. There are also many laws/regulations and federal program requirements that are intended to protect privacy. Some of these may specifically dictate what data can be made available, and may require certain safeguard methods to be used. Which methods you utilize is a decision left with your agency.

Online Resource

Health Information Privacy - De-identification

1. **Suppressing Fields/Records**

   Suppressing fields involves excluding, in their entirety, any fields (columns) in a dataset that contain the personally identifiable information or confidential data from the version made publically accessible. Suppressing records (rows) in a

Online Resource

Protecting the Confidentiality of Personally Identifiable Information

[6] Frequently Asked Questions—Disclosure Avoidance, Privacy Technical Assistance Center, May 2013, http://ptac.ed.gov/sites/default/files/FAQs_disclosure_avoidance.pdf

dataset involves excluding those rows in their entirety. However, you are encouraged to review other safe guard measures prior to suppressing records.

If any codes or identifiers are left in the dataset to allow records to be re-identified, they must not be known or accessible to unauthorized parties, or linkable to other external records.

2. **Suppressing Data Values**

Suppressing data values involve replacing data values with a generic term when only a portion of records in a specific field contain confidential data. For example, in our checkbook data, values for Vendor and Vendor Number have been redacted and replaced with "Restricted" where the record was flagged as confidential.

Online Resource

Privacy Technical Assistance Center, U.S. Department of Education

3. **Generalizing, Collapsing or Top/Bottom Coding Data**

Generalizing, collapsing or top/bottom coding data values involves using methods to make the data less precise as a means to ensure confidential data is protected. These methods can be used where unique characteristics result in small frequency counts.

Generalizing often applies to categorical or contextual data. This could include modifying a full address to present only the city, state and zip code, or providing a county name rather than a city or coordinate. There may also be instances where categorical data is too specific and can lead to individual identification, and needs to be replaced with a broader category (e.g. race categories for Chinese, Filipino, Japanese, Korean, and Vietnamese could be changed to a broader category of Asian). It may even be necessary to make your broader category "Other" in order to preserve as much detail in the dataset.

Collapsing or top/bottom coding involves transforming numeric data into ordinal[7] or interval[8] data. For instance, rather than providing the specific age of the individual associated with a record, you provide a range (e.g. 0 – 18 years old). Top coding and bottom coding involves collapsing top end and/or low end values into

[7] Ordinal data is similar to categorical data except that there is a clear ordering (e.g. low, medium, high)
[8] Interval data is similar to ordinal data except that there are equally spaced variables (e.g. 1 – 10, 11 – 20, … , 51 – 60, etc.)

one category, rather than transforming all values into ordinal or interval data. For instance, a high income could lead to the identification of an individual. Top coding would involve generalizing values above a certain level (e.g. >$500,000)[9]. The same principles apply for bottom coding – it is just on the opposite end of the range.

4. **Grouping Data**

Grouping data involves combining individual records (e.g. food assistance grants) into a single row of data with common characteristics using categorical data or combined groupings of categorical data (e.g. county and month), and totaling and/or averaging numeric values (e.g. grants awarded) contained in individuals records, and providing frequency counts of records (e.g. number of grants) within each group. By doing this, you are essentially publishing your dataset in aggregated form. Grouping data is most commonly used where the previous safe guards do not adequately protect the data for disclosure risk, or where you want to report on a confidential variable that may be disclosed in summary form.

However, there may still be instances where even aggregated data may still enable someone to derive information on or closely estimate information for specific individuals. In these instances, you may need to further suppress the data.[10] In instances where the numerical data needs to be further suppressed, consider whether rows containing the suppressed data could be further aggregated and reported to facilitate higher level summaries that couldn't happen if it were not present.

12.7 Explain Safeguards Used

If you employ safeguards to protect personally identifiable information and other confidential and sensitive data, you should describe these to ensure they are applied consistently over time (see Dataset Creation Steps – Confidential Data Redaction Steps).

13 Supplying Required Metadata

Your agency must supply metadata for datasets published on data.iowa.gov. Metadata describes the context, structure, and format of data – it is data about the data. It is a tool which we can use to manage our information resources. Much like you would use an index in the back of a book to find a topic of interest, metadata provides an "index" for information and data holdings that greatly simplifies searching. It also provides potential

[9] There are methods for basing top codes/bottom codes on a percentage of the total records, but this could result in the values falling into these categories to change over time.
[10] Two commonly used methods to determine the sensitivity associated with aggregated data are the (n,k) dominance rule and the (p,q) prior posterior rule.

users of the data a better understanding of the data – how it was collected, its purpose and use, and definitions for data it contains. Some general tips to writing metadata include:

- Avoid large words
- Limit or avoid use of jargon and acronyms
- Use present tense
- Use active voice
- Remember the general public should be able to understand what you have written

More information on and specific requirements pertaining to metadata can be found in Appendix B. Metadata.

14 File Formats

A number of file formats are acceptable for importing data into data.iowa.gov. The most common, however, will be a CSV file which is recommended for datasets without complex geometric structures. Data or attributes associated with lines or polygons will require the dataset to be formatted as a shapefile, a KML file or imported through a RESTful service.

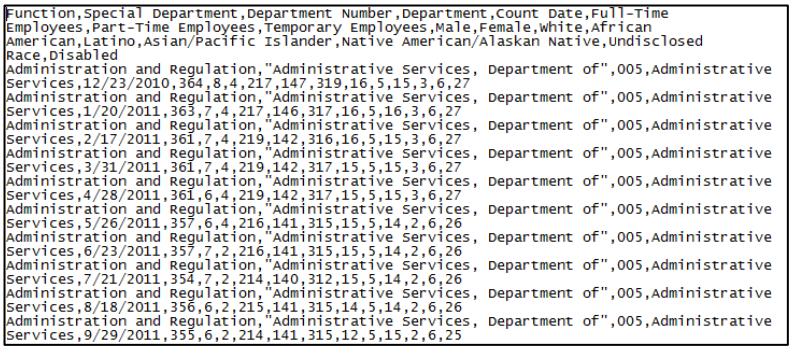


Figure 4 – Example CSV file format. One record per line with a hard return at end of line. Text is wrapped in this figure.

14.1 CSV file

A CSV file stores tabular data (both numbers and text) in plain-text form, see figure 4. So many programs and applications support some variation of CSV for exporting, which makes moving tabular data between programs with different and incompatible formats possible. Here are some basic characteristics of CSV files:

- There is typically one record per line (hard return after the record)
- Records divided into fields that are separated by delimiters (e.g. commas, semicolons, tabs). It is the individual fields within a record that become the dataset's columns when imported.
- Each record contained in the file should have an identical list of fields.
- The first row in the file should serve as your column titles or headings.

- Fields containing a line-break, double-quote, and/or commas should be quoted (i.e. containing a text qualifier) so the file can be processed correctly[11].

14.2 Shapefile

A shapefile is a digital vector storage format for storing geometric location and associated attribute information. It is a preferred format for more complex geographic structures, such as lines and polygons. While the name implies a single file, it is actually a set of several files. Your agency should organize the files into a zip file (.zip). At minimum, each shapefile (.zip) should contain the following:

- File defining the geometry (shapes) (e.g. geographic location of county boundaries) (.shp)
- File providing the attribute table (e.g. population and other demographic characteristic associated with each county) (.dpf)
- Projection file to ensure the feature locations are accurately rendered on the map (.prj)
- Shape indexing file for efficient processing (.shx)

Shapefiles should use the WGS-84 Geographic Coordinate System (EPSG/WKID: 4326) or Web Mercator (Auxiliary Sphere) Projected Coordinate System (EPSG/WKID:3857/102100).

14.3 Keyhole Markup Language

Keyhole Markup Language (KML) format specifies a set of features (place marks, images, polygons, 3D models, textual descriptions, etc.) for display on a map that is commonly associated with Google Maps and Google Earth. It and its zipped or compressed version, KMZ, can be used to import geospatial data into data.iowa.gov. Data should only provide a single geometric structure (e.g. point, line, or polygon) for a feature. Including multiple geometric structures inside the multigeometry tag is not supported by the system.

14.4 RESTful service

Agencies can also publish geospatial data in data.iowa.gov by providing the URL to the RESTful endpoint for a map or individual map layer on an ArcGIS Server version 10.0 or above. Maps published in this manner can have points, lines and polygons in different layers. Map layers contained in a feature service need to be entered individually, but can be combined into a single map later on. This approach allows maps to have customized point, line and boundary color. The map is controlled and updated on your own server – not in data.iowa.gov. The system, data.iowa.gov, will call every time a user loads the page. Agencies need to:

[11] If you are saving an Excel file as CSV, Excel will automatically put quotes around text fields requiring them.

- Set up your services to reproject on the fly into WGS-84/Web Mercator, which is one of the features of ArcGIS 10.x.
- Ensure your ArcGIS server runs on an https SSL cert. If not, most browsers will reject the non-authenticated content.

15 Selecting a License

Data published on data.iowa.gov should be freely used, reused and redistributed by anyone with no or minimal restrictions to be considered open. State agencies should select and apply a license to your data to provide clarity on how the data may be used. Below are licensing options:

1. Public Domain – Indicates the dataset is not subject to copyright protection. As such, the data can be copied, modified, and distributed, even for commercial purposes, without asking permission.
2. Creative Commons – Provides some baseline rights regarding attribution, distribution, commercial use, and derivative works of copyrighted data. More information regarding license options can be found at creativecommons.org – about the licenses.
3. Open Data Commons Public Domain Dedication and License – Places the data in the public domain waiving all rights. More information can be found at opendatacommons.org – Open Data Commons Public Domain Dedication and License (PDDL).
4. Open Database License – Allows users to freely share, modify, and use a database while maintaining this same freedom for others. It governs the database itself and not the contents of the database individually. More information can be found at opendatacommons.org – Open Database License (ODbL).

16 Importing New Datasets

Agency data coordinators and data stewards are able to create new datasets on data.iowa.gov through a web-based interface. Each account in the system has a dashboard, which you are automatically directed to once you sign in. If you are not, you can easily navigate to your dashboard by clicking the "Hello" link above the main search box for the website. Once you are on your dashboard, click the "Create a New Dataset" button to get started.

Video Tutorial

Import Datasets

16.1 Import Options

There are different options for importing available, see figure 5 below.

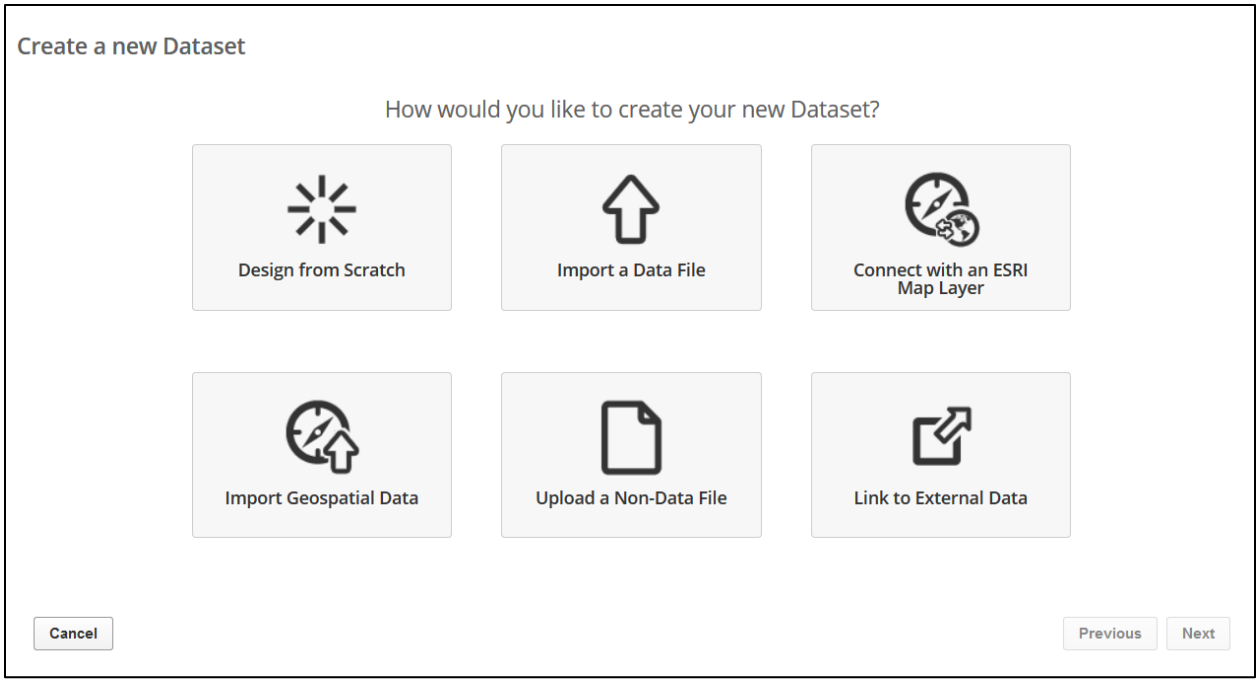


Figure 5 – Create a New Dataset. Users of data.iowa.gov have different options available for creating datasets.

- "Design from Scratch" allows you to create the dataset schema without importing the data. This should be used where the data files are more than 250,000 rows or 150 MB, and/or are going to be frequently updated, and will imported and updated automatically, see Automating Updates.
- "Import a Data File" allows you to import data (where file is less than 250,000 rows or 150 MB) and create the dataset schema. *Most agencies will use this option.* When you select this option you will be prompted to identify the location of the dataset – whether it is on your computer or on the internet.
- "Connect with an ESRI Map Layer" are for those agencies with RESTful service described previously.
- "Import Geospatial Data" is used to import a Shapefile or Keyhole Markup Language file.
- "Upload a Non-Data File" and "Link to External Data" will generally not be used. However, there may be instances where it makes the most sense. Most notably would be linking to external data to catalog public online databases. Agencies should consult with the state data administrator if you believe this choice would be the best option for you.

16.2 Reviewing Data File Schema

The system makes an educated guess on the data type based data contained in each column. However, you are able to change it if it is not correct. The system supports variety of data types which can be found in Appendix C. Data Types. Although it is possible to

change the data types later on, it can be difficult if the dataset is large (e.g. > 50,000 rows) and can result in loss of data. It is best to review and correct when setting up the dataset originally.

16.3 Creating Location Columns

When importing your data, you will need to create a location column with any geospatial data you may have within the CSV file, such as address columns and latitudes and longitudes.

Help Article

Input geospatial data using the location column

16.4 Provide Metadata and Set Permissions

Once you review the data schema and create any applicable locations columns, you will be presented a web form contains the fields highlighted in Appendix B. Metadata. However, not all metadata is available to you on the initial import. You are encouraged to edit your metadata (see Editing Metadata) later on to provide all the necessary information to ensure the public understands what your data is all about.

On this form, you also have the opportunity to indicate if your dataset is public or private. Data published to data.iowa.gov should be marked as "public" when importing your data. It is important to note that even though it is "public" your new dataset will not be accessible to the public until it has been published.

16.5 Changing Column Properties

To change column properties, simply move your mouse arrow to the column you wish the change. On the right side of the column, you will see a menu icon, click on it, see figure 6. A drop box will appear, select the "Edit Column Properties" option. This will open the Column Properties panel where you can rename the column (i.e. provide a label for the column), provide a description for the column, and change the column's formatting.

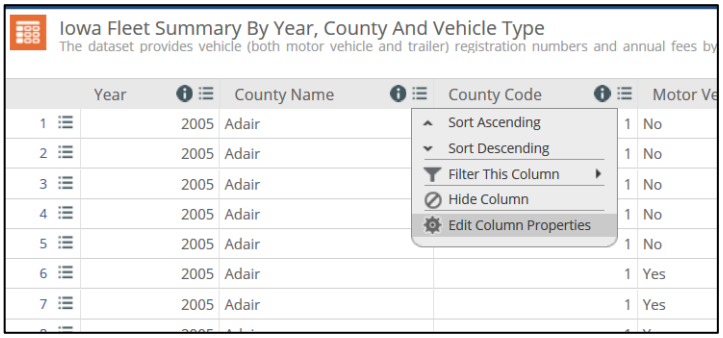


Figure 6 – Column Menu & Edit Column Properties. To edit column properties, click the menu icon (three lines) next in column heading area.

Names for the columns should clearly describe the columns contents. The description should include further detail about the column, including any commonly used abbreviations. For numerical data, the description should also include how it was measured (i.e. method or methods used), if applicable, and provide the units associated with the measure. This will help ensure that those who use the dataset have a good understanding of the data contained within it.

16.6 Publishing Datasets

New datasets are working copies that are not accessible to the public until you publish it. This will allow you to share datasets (see Sharing Datasets) for review and collaboration until it is ready to be made public. Before you can make your dataset public, you'll need to publish your working copy to make a published copy of the dataset for the public to access. To publish your dataset, just click "Publish Dataset" button adjacent to the dataset title.

17 Sharing Datasets and Changing Owners

It is a good practice to share your datasets with other users who will be involved with updating, reviewing and/or approving your data. You may do this by clicking "Manage" on your dataset. In the manage panel, you click "Sharing" then the "Share this Dataset" button. You can then enter emails of the individuals you would like to share the data with and specify appropriate permissions. There are three different permissions you may give to those users you share your dataset with:

- Viewer – allows users to view a working copy of the dataset, view a private dataset, and have quick access to the published dataset in their dashboard - "Shared to Me" tab.
- Contributor – allows users to not only view the dataset, but edit it as well. However, they are not able to publish changes.
- Owner – allows users to view, edit and publish your data, as well as, make changes to the dataset schema.

You may also transfer ownership of your dataset to another user via the manage panel. Just click "Ownership" and enter the email of the user you want to transfer ownership of the dataset to. You will be required to select from list of options. Please note that they must have an account on data.iowa.gov, see Data Ownership

Your agency retains ownership over the datasets that you publish. All data and datasets remain the property of the originating Agency and public users acquire no ownership rights to your agency data or datasets. The datasets published on data.iowa.gov or any Iowa transparency website becomes a public resource available to anyone with access to the Internet. The public use of the datasets may include development of applications. In this case, the developers retain all intellectual property ownership in their applications, excluding the agency data itself, whose ownership continues to reside with the agency.

17.1 Responsibility for Quality, Integrity and Security

Your agency is responsible for all aspects of the quality, integrity, and security of dataset content published. You do not relinquish control of your data when a dataset is published on data.iowa.gov or any Iowa transparency website. Your agency is responsible for ensuring that all of your published data has been reviewed by appropriate management for confidentiality, privacy, security and all other content limitation issues before the data is published.

System Accounts.

18 Maintaining Datasets

Datasets published on data.iowa.gov must be maintained to ensure they remain accurate and accessible and that the data is current.

18.1 Update Frequency

Agency Data Stewards are responsible for identifying an update frequency for each of your agency's published datasets, as part of your datasets' metadata. The frequency set should be consistent with operational updates to the data. Both Agency Data Coordinators and Agency Data Stewards are then responsible for ensuring that your content updates are maintained and published according to the identified frequency.

Video Tutorial

Edit Datasets

18.2 Content Updates

To edit datasets via the web-based interface, you will need to open your dataset, and click "Edit" button to open the edit panel. On the edit panel, you will need to click "Edit Dataset" to create a working copy of your dataset.

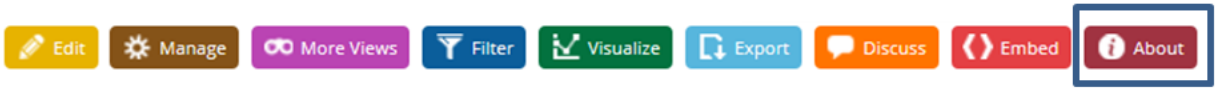
A working copy is an editable version of your dataset that you can change and collaborate on with your coworkers. Changes made to the working copy are not reflected on the published copy until you are ready. When you're done with making your changes, you can "publish" the working copy to be the new published copy, and your changes will become accessible to the public.

Help Article

Append or Replace Dataset Rows

In the working copy, you can make changes to individual cells, or launch the Append and Replace wizard to add rows to your dataset, or replace your dataset entirely.

19 Editing Metadata



Not all of the metadata fields will be available to you to complete as part of the initial import process. You can access the metadata form by clicking the "About" button, then "Edit Metadata" link. The web form contains the fields described in Appendix B. Metadata. Please review the appendix and complete the required information.

Help Article

Edit Dataset Metadata

You may also edit the metadata about your dataset at any time without creating a working copy of your dataset.

20 Automating Updates

Agencies can schedule automated imports/updates of their data. This is particularly useful where you intend to update your data frequently (i.e. daily or weekly) and/or your dataset is large (e.g. > 250,000 rows or 150 MB). If you are interested in automated updates, you will need to provide the following information to the state data administrator:

- Name of the dataset, and its URL
- Name and email of the individual responsible for updates
- Update frequency (i.e. daily, weekly, monthly)
- Indicate whether update files will append to or replace existing data
- Describe columns used to populate the location column where applicable.

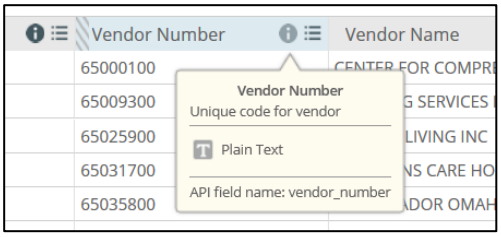


Figure 7 – API Field Name. The API field name can be found by hovering mouse over column information icon.

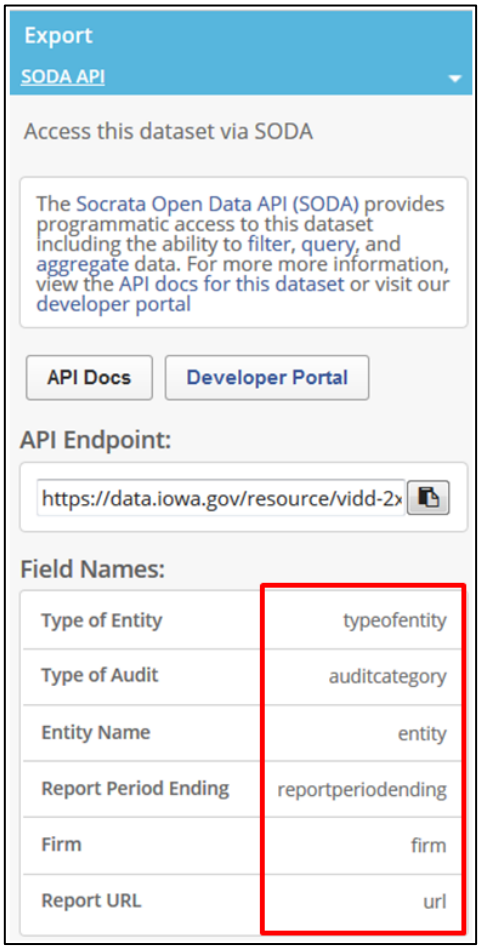


Figure 8 – API Field Names List. To quickly access a full list of the API field names in the dataset, click "Export" then "SODA API", the complete list is shown under Field Names heading on the right side (highlighted). Column labels are to the left.

Files should be named using their dataset id[12] (e.g. b3t9-awkp.csv), and be saved as a CSV file. Header information in file must contain the API field names of each column in the dataset. If not known, the API field names can be found by hovering over the information icon for each column, as shown in figure 7, or opening the Export panel, as shown in figure 8. Setting your dataset up with Joins

Many datasets will be created with the use of complex queries involving joins of two or more tables. Joins have the potential to affect the number of rows returned and ultimately the accuracy of the dataset. To help avoid problems, it is a good practice to begin joins with the most granular table or query. Often times, this table or query will contain the key Numeric Data for your dataset that you want citizens to be able to summarize (e.g. count of offenders in prison, sum of payments made, etc.). By beginning

[12] The dataset id is the alphanumeric sequence – four characters, a dash, and four additional characters found at the end of the dataset's URL (e.g. https://data.iowa.gov/Health/Monthly-Medicaid-Payments-By-Vendor/b3t9-awkp)

joins on this table or query, you can better trust that any aggregates on your primary table or query will be accurate. Oftentimes, the tables or queries you join to the primary table or query will contain Contextual Data or Categorical Data (e.g. institution, payee, etc.), which facilitates summarizing the data.

Problems can occur if one row in your primary table or query can match up with multiple rows in a joined table or query. This can lead to duplicate numeric values, which would cause inaccuracies with any aggregate function.

Column Identifiers at the start will make this process easier. The state data administrator or the Office of the Chief Information Officer will provide the network location/ip address to place your update file.

The system will check daily for update files. Once the files have been imported, they will be placed in the "complete" folder. If the import failed, the file will be moved to the "failed" folder and the system will generate a text file highlighting the reason the import failed.

21 Preparing Data to View

With very detailed, granular data, it is often necessary to take steps to better summarize your data before Creating Data Views. This can include filtering your data to focus on a specific subset, grouping your data, and/or applying conditional formatting.

Video Tutorial

Basic Filtering

21.1 Filter Data

In the filter panel, you are given the option to filter a dataset based on its contents. Once you have opened the filter panel, and clicked "Filter" you can click the "Add New Filter Condition" button. When creating a filter condition, you must select the column you would like to filter by and the operation you would like to use, and the condition you would like to apply (e.g. a word, number, date, or range). Table 1 highlights filter operations available, and the data types they can be used for.

Filter Operation	Text	Date	Numeric
is	✓	✓	✓
is not	✓	✓	✓
starts with	✓		
contains	✓		
does not contain	✓		
is before		✓	
is after		✓	
is less than			✓
is at most			✓
is greater than			✓
is at least			✓
is between		✓	✓
is blank	✓	✓	✓

Table 1 – Filter options and data types.

21.2 Grouping Data

In the filter panel for a dataset, you are also given the option to roll-up data. Roll-ups allow you to group and summarize Numeric Data or counts. It is often necessary that you "roll-up" your data in order to create a meaningful chart. Roll-ups are created by selecting one or more columns containing Categorical Data to group by, then selecting one or more columns containing numeric data and a function to apply (e.g. average, sum, maximum and minimum). You can also count records, which can be applied to columns containing either numeric values or text.

21.3 Conditional Formatting

Conditional formatting, found in the filter panel of a dataset, can enhance your charts and maps. It allows you to create rules with conditions, and define what formatting to apply, such as setting a color or specifying an icon to use, when those rules are met.

Help Articles

Grouping and Roll-ups

Use conditional formatting

22 Creating Data Views

Poring over records in a dataset is a slow and tedious process for identifying relationships, patterns and trends in data (if they can be identified at all). Most people are not going to take the time to do it. Data views, such as charts and maps, offer a way to make data presentation interesting, aesthetically pleasing and hopefully informative. But most importantly, good charts and maps help reveal patterns, trends and relationships in the data that would have gone unnoticed by even those most familiar with the data.

22.1 Charts Types

There are a number of options, see figure 9, available to chart your data. What you select will depend on what you need to show:

- Comparison among categories
  - Column Chart
  - Bar Chart
- Comparison over time
  - Line Chart
  - Timeline Chart

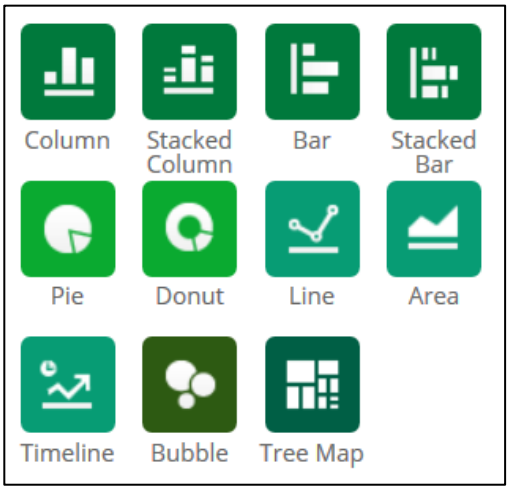


Figure 9 – Chart Selection Options. Users of data.iowa.gov have a number of charting options available for use.

    - Area Chart
- Composition – multiple periods
    - Stacked Column Chart
    - Stacked Bar Chart
- Composition – single period
    - Pie Chart
    - Donut Chart
    - Tree Map
- Relationships
    - Bubble Chart

22.2 Setting Labels

Once you select the chart type, you are able to select a column with Categorical Data to use as labels. For column charts, bar charts, line charts, area charts, bubble charts, and timeline charts, labels will determine placement of values along x-axis (horizontal axis along the bottom, except for bar charts where it is vertical and along the right side). However, for pie charts, donut charts and tree maps, it establishes the data series.

22.3 Setting Values

You can also select one or more column with Numeric Data to use as values. For column charts, bar charts, line charts, area charts, bubble charts, and timeline charts, values dictate placement of values along the y-axis (vertical axis along the right, except for bar charts where it is vertical and displayed along the top). If more than one column is selected, each one is presented as a separate data series. However, for pie charts, donut charts and tree maps, you are limited to one value and it determines the size of the slice or box.

22.4 Advanced Settings

Some chart types allow additional grouping using columns containing categorical data. This typically creates new data series (e.g. cluster of bars/columns, multiple lines in a line chart, and individual segments of a stacked column or bar).

22.5 Customizing your Chart

You also have a variety of options for customizing how your chart is presented, such as:

- Picking colors used for data series
- Indicating whether you want values and labels displayed on the chart itself
- Adding annotations, such as targets, limits or baselines.
- Adding and customizing a legend (if you only have one data series, don't include a legend)
- Customizing flyouts that are viewable when a cursor is placed over a point, bar, bubble, slice or box.

- Providing axis titles and customizing the axis itself.

22.6 Maps

Maps give you the ability to display where points of interest (e.g. a school building, county park, permitted facility, etc.) are located relative to one another. If you have locational data (i.e. geographic coordinates, or a geocoded address), you can create point, heat and boundary maps on data.iowa.gov with relative ease.

Help Articles

Make point, heat, and boundary maps

Create a calendar

22.7 Calendars

If your dataset has one or more date/time columns, you have the ability to display events on a monthly calendar view.

22.8 Saving your View

If you are creating a view from a dataset, you will be given options:

- Save – applies your changes to that particular view (does not apply when creating a view from the underlying dataset)
- Save as – creates a filtered view which is static instance of the underlying dataset or view, reflecting the changes applied. This view can be linked to separately. The underlying dataset or view will remain the same.
- Revert – undoes any applied changed and will take you to the saved state of the underlying view.

22.9 Naming your View and Other Metadata

When using "Save as" to save your view, you are asked to name the view. The name you provide should be in plain English and include sufficient detail to facilitate search and discovery. Some basic elements should be considered when coming up with a title for your view:

- The main numeric data being summarized should provide the foundation for your title (e.g. Total Assessed Property Values)
- Known timeframes your view is limited to should also be used if applicable (e.g. 2012 Total Assessed Property Values)
- Groupings used to summarize underlying data (e.g. 2012 Total Assessed Property Values by County)

The metadata from the underlying dataset is the default metadata for your view, but can be edited so that the description and other fields are more specific to the view itself. You can edit the view's metadata (see Editing Metadata) the same way you do for your datasets.

23 Embedding a Dataset or Visualization

Using data.iowa.gov, you are able to embed any dataset, chart, map, calendar, or filtered view on to your website. This provides additional ways for you to share your data with others. By embedding content, anytime the underlying data is updated, the embedded dataset, chart, map, calendar or filtered view is also updated. No longer do you have to update content in multiple places. To embed content, navigate to the dataset or view of interest, and follow the steps below:

- Open the embed panel by clicking the "Embed" button
- Select the size you want
- Copy the html code provided
- Place the html code in the website page you want the dataset, chart, map, calendar or filtered view to display

Help Article

How to Embed a Dataset or Visualization

24 Additional Help & Training

The open data portal, data.iowa.gov, has a lot of functionality – only a portion of which was covered in this handbook. You are encouraged to make use of available help and training resources noted below.

24.1 Socrata University

Our vendor, Socrata, offers a webinar series to educate new users with many of the features available in data.iowa.gov. Two foundation courses that are routinely offered are:

- Socrata 101 – This webinar provides basic site training, publishing, built-in dataset features, data visualization, embedding, and more.
- Socrata 201 – In this webinar, you will learn how to create more advanced visualizations, such as stacked column and bubble charts, as well as point, heat and boundary maps. Conditional formatting and map mashups are also covered.

You can register for web courses on the Socrata University webpage.

24.2 Open Data Portal – Knowledge Base

Our vendor, Socrata, also provides a support website which contains numerous articles and videos related to the open data portal. You may find the information you are looking for on the Open Data Portal – Knowledge Base.

24.3 Individual Assistance

Please contact the state data administrator for individual assistance.

25 Conclusion

Congratulations on taking the necessary steps to make your data open and available to the public. As state agencies begin to publish more and more data on data.iowa.gov, we will begin to realize the true benefits of open data:

- The cost and purpose of government services are better understood
- Barriers are broken and data is frequently leveraged for new insights
- Citizens are more actively engaged in their government
- Decisions are data driven
- The re-use of data creates economic opportunities

Appendix A. Drafting Your Agency Data Plan

This appendix walks you through drafting your agency data plan, which includes compiling an inventory of key agency data, conducting impact and difficulty assessments, prioritizing your data for publication and outlining a schedule. A template is available at https://data.iowa.gov/policies.

Data Inventory

Finding out what data your agency collects, maintains or holds – even if it is historical data – is an excellent first step towards identifying datasets to publish on data.iowa.gov. Below are some activities to help get you started:

- **Review the data already posted on your websites.** Look at Microsoft Excel files or public facing applications, which allow visitors to search for records. The data may not necessarily be accessible in bulk, or available through machine-readable mechanisms, but can serve as a good starting point.
- **Review published reports.** Published reports are often populated with data which is compiled or aggregated from internal data systems. For example, a public report may indicate that an agency has closed 100 projects in the last month. The internal data system, which maintains information for each project, will likely have additional details that can be made public, such as, the type of project, its location, etc.

> **Tip**
>
> Google search bar can facilitate finding data files on your website by entering the URL for your website and file type (e.g. site:www.educateiowa.gov filetype:xls).

- **Review trend and statistical analyses performed.** These analyses often use data from various sources, and are typically associated with issues or problems your agency/institution deems important. Your constituency may also find this information of value.
- **Review reports provided to federal agencies or the State Legislature.** These reports (and their underlying data sources) can help identify data which can also be provided to the public. In addition, meeting these reporting requirements (particularly statutory ones) might be accomplished simply by making the dataset(s) available on data.iowa.gov.

As you are reviewing data that your agency or institution collects, maintains or holds, you should document any data that would help explain your mission/operations, detail key results achieved or issues impacting your agency, and/or show progress on strategic initiatives into an inventory. Below are some variables you will want to capture pertaining to data identified:

- Title of the dataset
- A brief description of the relevant data or information contained in the dataset
- The data source
- The date data collection started and stopped (if applicable)
- Indicate whether the data is subject to copyright protection
- The URL where the data is available (if published online)
- The dataset ID[13] (e.g. b3t9-awkp) where already published on data.iowa.gov
- The frequency in which the data is updated (if applicable)

Impact Assessment

Once the review is complete, your agency/institution should assess your data's impact. The more valued a dataset is, the higher the impact it is likely to have once published. The following questions are to help guide you in identifying high value data:

- **Does the data show progress on strategic initiatives?** Data related to the Governor or your agency's strategic initiatives is of value if it can help demonstrate progress being made and/or if your agency's efforts are having the desired effect.
- **Does the data help tell your agency's story?** If the data helps improve the public's understanding of your agency's mission and operations and/or quantifies results achieved by your agency, it is of high value.
- **Does the data provide insights on key issues affecting your agency?** Data which helps explain issues or answer questions can be of great value to your agency and its constituents.
- **Is the data frequently requested?** As demand is known and quantifiable, this should raise the value of data for publication. If the dataset is requested on a recurring basis, then your agency may reduce duplication and obtain efficiencies by publishing data on data.iowa.gov.
- **Does the data have a direct impact on the public?** The data is likely of higher value if it is already apparent there is a deep impact and interest by the public.
- **Is the data in strong demand with your constituencies?** The data might be of higher value to specific, narrow interest groups which may be your agency's core

High impact means data:

- ✓ Tracks strategic initiatives
- ✓ Tells agency story
- ✓ Is frequently requested
- ✓ Has a public impact
- ✓ Is in strong demand
- ✓ Is of timely interest
- ✓ Costs high $$$ to collect
- ✓ Provides an economic opportunity
- ✓ Facilitates reporting
- ✓ Encourages cross-agency collaboration

**Data does not have to meet all conditions to be considered high impact.

[13] The dataset id is the alphanumeric sequence – four characters, a dash, and four additional characters found at the end of the dataset's URL (e.g. https://data.iowa.gov/Health/Monthly-Medicaid-Payments-By-Vendor/b3t9-awkp)

constituency for those issues. It is important to not overlook these constituencies even if demand from the public overall is low.

- **Is the data of timely interest?** Announcements of progress or success or reactions to public criticism can be strongly supported by publishing related data, should it exist.
- **How much does the data cost your agency to collect and maintain?** If you spend a great deal of money on a particular set of data, then it is highly likely that others would like to access it.
- **Could availability of the data create an economic opportunity?** In many cases, this will be unknown to your agency in advance. Some of the greatest successes with making public data available have involved government data being commercially appropriated in useful ways. Any anticipated commercial use of your data should be taken into account.
- **Can publication of the data help satisfy regulatory, statutory, or grant reporting requirements?** Some required reports do not require extensive narration, and may be satisfied by publishing datasets alone, or in combination with interactive reports/summaries.
- **Can publication of the data facilitate collaboration with other state agencies?** Certain state functions may involve multiple agencies requiring access to similar data. Your agency may collect data that is of considerable value to another agency.

Difficulty Assessment

Your agency should also assess the difficulty related to publishing a dataset, which is most often tied to how easily the data can be extracted and the quality of the data. The following questions are intended to help guide your assessment:

- **Is your data structured?** If your data is contained in a fixed field with in a record or file (e.g. contained in a database or spreadsheet), it will be much easier to publish compared to data in a document or paper format. Structured data typically has a data model that defines the fields data will be stored in and specifies how the data will be stored (e.g. data type – numeric, date, text, etc.). Structured data is much easier to extract from its data source(s) – thus making it easier to publish.
- **Is your data complete?** Missing or incomplete data prevents users from being able to effectively aggregate and compare values. If your dataset is missing relevant

Low Difficulty means data is:

- ✓ Structured
- ✓ Complete
- ✓ Unambiguous
- ✓ Consistent
- ✓ Non-redundant
- ✓ Based on standard references/protocols
- ✓ Free of confidential/sensitive data

**Data does not have to meet all conditions to be considered low difficulty.

data values, it may be necessary to complete records from other sources (e.g. paper records or electronic documents). The more extensive or widespread the gaps in your data are, the more difficult it may be to publish your dataset.

- **Is your data unambiguous?** Data ambiguity arises when what the data represents is not precisely defined. This can lead to data values being misinterpreted. For example: DHS can represent two different government agencies: Department of Homeland Security at the federal level or Department of Human Services at the state level, or a reference to a person (e.g. Smith, J.) could be associated with different individuals. Having to correct ambiguity in your data will make it more challenging to publish.
- **Is your data consistent?** Data inconsistency occurs where data values refer to the same thing but are recorded differently. For example, Mt. Pleasant and Mount Pleasant refer to the same town in Iowa, and Dept. of Public Health, DPH and Iowa Department of Public Health all refer to the same state agency. However, since they are not recorded in a consistent way, values cannot be properly aggregated and compared. Correcting data inconsistencies can make publishing your data more challenging.
- **Is your data redundant?** Redundant data usually occurs where data values for the same thing are recorded in multiple places. This could potentially lead to contradictions in the data. For example, if vendor addresses are entered on individual financial records, rather than in a unique vendor record, there is the potential for different addresses being recorded. When this happens, data users would not know which address is the correct one. Having to determine which data is correct will make publishing your data far more challenging.
- **Is your measured data based on a standard?** Measured data lacking a standard reference method or measurement protocol lends itself to uncertainty, as it cannot be easily replicated. Additionally, if measured data lacked a standard and was collected by multiple individuals – the accuracy of the data becomes questionable. This is perhaps the most difficult data quality issue to deal with.
- **Is your data confidential or sensitive?** If your dataset contains confidential or sensitive data (e.g. data protected by state law, such as Iowa Code Section 22.7 or other applicable Iowa Code section, or federal law, such as the Health Insurance Portability and Accountability Act, Social Security Number Protection Act, and Family Educational Rights and Privacy Act), it will be more difficult to prepare for publication. De-identification and other disclosure requirements can greatly increase the burden of publishing the data for public use if protocols and procedures have not been developed. Confidential or sensitive data in some cases could prevent a dataset from being published as an open dataset altogether.
- **Are there existing processes in place for publishing your data?** Your agency may be able to leverage existing processes to publish the data, such as exports for periodic department reviews, or routine exchanges of data with other agencies (e.g. data sharing agreements). It would also include any quality assurance processes to verify the quality and integrity of your datasets. Having such existing procedures in place may make the data easier to publish.

Setting Priorities

Once your data inventory is complete and the data has been assessed, you should prioritize your data. It is recommended that you use an impact/difficulty matrix, as shown in Figure 10, to help. You will need to score each dataset identified that is not already published on data.iowa.gov. Scoring should be based on the Impact Assessment and Difficulty Assessment previously discussed, and can be done on the template provided by the Department of Management.

For impact, each dataset should be given a score between 1 and 10, with 1 representing low value and 10 representing high value. The same is done for difficulty with 1 representing data that is very easy to publish and 10 representing data that is very difficult to publish. If multiple raters are used, the dataset's score should be the average individual rater score.

Once a score has been given to each dataset, the dataset's scores should be compared to the average impact and difficulty score for all datasets to determine whether a dataset is high or low for impact and difficulty. The template provided by the Department of Management will calculate the results. High Impact, Low Difficulty datasets should be given the highest priority.

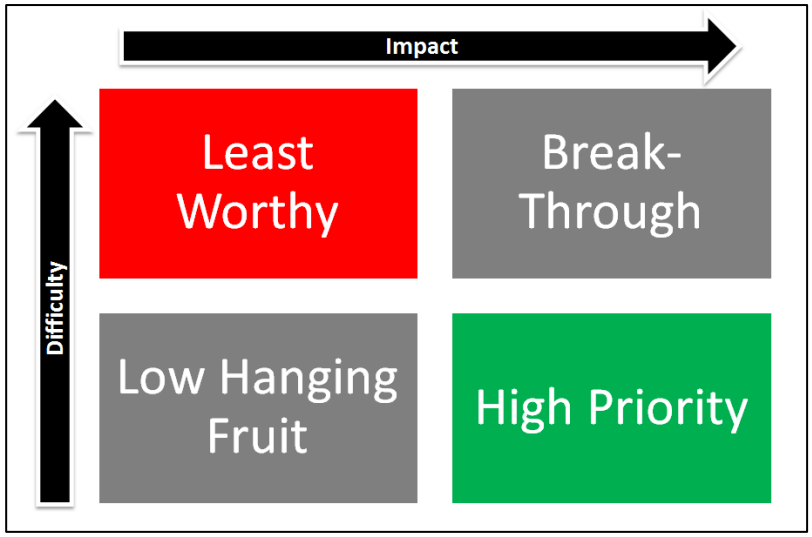


Figure 10 – Example of the impact/difficulty matrix. High Impact/Low Difficulty datasets should have the highest priority.

Schedule

Once datasets have been prioritized, you will need to set reasonable target dates and identify responsible data editors(s) for those datasets your agency intends to publish on data.iowa.gov. Your plan should contain target dates for the upcoming fiscal year.

Appendix B. Metadata

Establishing a common metadata vocabulary is critical to effective communication and to allow us to share our information with others. The following is a list of metadata elements provided and descriptions of what they should include.

General Information

Dataset Title	Create a title for the dataset. It should be in plain English and include sufficient detail to facilitate search and discovery. Some basic elements should be considered when coming up with a title for your dataset: • The main numeric data available within your dataset should provide the foundation for your title (e.g. Vendor Payments; Assessed Property Values; Local Option Sales Tax Rates & Payments) • Known timeframes your dataset is limited to should also be used if applicable (e.g. FY 2013 Vendor Payments; 2012 Assessed Property Values) • Groupings used to summarize underlying data where record level detail is either not available or not provided due to sensitive or confidential data (e.g. 2012 Assessed Property Values by Tax District; FY 2014 Monthly Medicaid Payments by Vendor)
Description	Highlight information about what the dataset contains, and why is it important. The description needs to provide sufficient detail to enable a user to quickly understand whether the dataset is of interest. You will want to ensure your description is easily understood – appropriate to the public's reading skills, and knowledge. It should also be clear and direct, free of unnecessary jargon, acronyms and abbreviations. Oftentimes acronyms and abbreviations have multiple meanings in different areas of government, industry, or even walks of life. As such, unintended meanings for abbreviations and acronyms can cause confusion and uncertainty in what your data conveys.
Category	Select the main thematic category for the dataset. The following options are available to select from: • Communities & People - Data about the characteristics of Iowa communities, our people, and how we live together • Economy - Data about economic activities, employment, agriculture, and business and industry in Iowa • Education - Data about student achievement, and elementary, secondary and post-secondary education in Iowa • Environment - Data about Iowa's landscape and habitat as well as the protection and conservation of our natural resources • Government - Data about government, and its spending, taxes, performance and operations • Health - Data about factors affecting health, health conditions and services available to Iowans • Transportation - Data about the conveyance of people and goods across Iowa via roads, airports, waterways and trails • Utilities – Data about Iowa's energy, water and waste systems and communications infrastructure

Tags/Keywords	Enter keywords that help users discover the dataset. Include terms that would be used by both technical and non-technical users. Each term should be separated by a comma. All terms will be converted so characters are lower case.

Licensing and Attribution

License Type	Select a license that is applicable to the dataset. You have the following options available: • Creative Commons • Italian Open Data License 2.0 • Open Data License • Public Domain
Data Provided by	Cite the agency, division, bureau and/or program as well as database, survey, report or related resource (where applicable) from which the described dataset is derived (e.g. Iowa Department of Administrative Services, State Accounting Enterprise, I3 Data Warehouse).
Source Link	Where available, provide the publically accessible web address for the database, survey, report or related resource from which the described dataset is derived.

API Endpoint

Row Identifier	Select column containing permanent identifier for rows in dataset. This gives developers a level of comfort knowing that they can use these columns to power their application. Even if other columns get deleted or added, they are ensured that the applications built off of key identifying information within the dataset (for example an ID number for each row) will not break.

Thumbnail Image

Thumbnail Image	A public domain image or one that your agency has the copyrights to that represents the data contained in the dataset and can be used on data.iowa.gov's homepage to feature the dataset. It should be cropped so that it is a square.

Attachments

Attachments	Related documents such as data dictionary, quality assurance/quality control documentation, technical information about the resource, developer documentation, etc.

Contact Information

Contact Email	Email for the contact responsible for answering questions related to and receiving feedback about the dataset. Address will not be displayed publically, and will default to the account's email if left blank. Unless agencies have a specific organizational email to use for this purpose, it is recommended that the field be left blank and default to the email on the account.

Coverage

Time	The date or time interval applicable to the dataset. It is important to provide this so old data is not presumed to be current. If you require time details, please contact the Department of Management. ***Date Representation*** A date represents a time period. Agencies can represent dates using either calendar dates or week dates. A calendar date is represented by the format YYYY-MM-DD. YYYY is the year in the Gregorian calendar (e.g. 2014), MM is the month of the year falling between 01 (i.e. January) and 12 (i.e. December), and DD is the day of the month falling between 01 and 31. A calendar date can be shorted to only reflect the month (e.g. YYYY-MM) or year (e.g. YYYY). A week date is represented by the format YYYY-Www-DD. Again YYYY is the year in the Gregorian calendar, W indicates weeks, ww is week of the year falling between 01 and 52, and DD is the day of the week with Monday being 01 and Sunday being 07. ***Duration Representation*** Duration is a component of time (e.g. one month or one year), and is represented by the format P[n]Y[n]M[n]D or P[n]W. In this format, P indicates a time period, and [n] represents the number of components. The Y, M, and D represent years, months, and days respectively. W represents weeks. (e.g. P2Y = two years, P5M = five months, P2Y6M = two years, six months, P10W = ten weeks, etc.). ***Using a single date*** A single date (e.g. 2014-06-30; 2014-06; 2014) should be used where the data was captured or is applicable to a single period in time, and no further updates are planned. ***Using a time interval*** A time interval is the intervening time between two dates represented by <start date>/<end date> (e.g. 2013-07-01/2014-06-30). It should be used where data being published was collected over a period of time, and no further updates are planned. ***Using a repeating time interval or duration*** Repeating time intervals or duration should be used where the dataset will be updated on some periodic repeating interval. Typically these are represented by R[n]/<start date>/<duration> or R/<duration>. R indicates a repeating interval, and [n] represents the number of repetitions (e.g. R12/2013-01/P1M – Data is updated monthly 12 times starting January 2013). If the [n] is not included, the number of repetitions is unbounded (i.e. endless) (e.g. R/2010/P1Y – means data is updated annually starting in 2010 and continuing through present).
Area	Provides the name of the geographic place of which the dataset is related. Agencies should use names of geographic features provided in the USGS Geographic Names Information System (e.g. Iowa; Muscatine County, IA; Des Moines, IA; Rathbun Lake, Appanoose County, IA; Big Creek State Park, Polk County, IA; Walnut Township, Madison County, IA). Multiple geographic places can be entered, where appropriate, and should be separated by a semicolon.

Disclaimers

Completeness	Provides information related to missing or incomplete data that would prevent users from being able to effectively aggregate and compare values. This could be either the result of data quality issues, or due to the need to protect confidential data. Agencies should also highlight any items that the public may perceive to be in the data (e.g. State of Iowa Expenditures do not include expenditures made by Regents institutions).
Limitations	Provides information related to any limitations on how the data can be used and/or summarized. For instance, aggregate monthly data providing the number of unique recipients cannot be totaled to determine the number of unique recipients over a year, as one recipient may be included in multiple months.

Updates

Agency	Designates the state agency or institution that owns the dataset and is responsible for updates.
Update Frequency	Designates the frequency associated with data updates (to be associated with the update time period below). Options include: As available (noting updates are random, and not on a set frequency), Every, and options for Every 2 through Every 30.
Update Time Period	Designates the time period associated with data updates (related to the frequency noted above). Options include: Day, Days, Week, Weeks, Month, Months, Year, Years
Update Notes	Provides information on when new data is typically available. (i.e. For datasets updated every three months, agencies can list months data updates will be published. For annual updates, agencies can list the month of the year new updates are typically available.)

Dataset Creation Steps

Data Export Steps	Field is not publically viewable. Used to highlight the source(s) for the data, name the queries or procedure run to produce the data extract, and any steps taken to transform the data to make it available for public consumption. This documentation is intended to help future owners of the dataset ensure the dataset continues to be maintained in a consistent manner.
Quality Assurance Process	Field is not publically viewable. Highlights data checking and review steps used. Summarizes checks used and steps taken to review your data, and clean or correct data where appropriate.
Confidential Data Redaction Steps	Field is not publically viewable. Highlights steps taken to de-identify or redact data that is confidential or sensitive. Should state any disclosure thresholds that are used.

Appendix C. Data Types

The following table provides a list of data types available in data.iowa.gov. The last four data types listed are not available when creating/updating a dataset using a file. You may add columns using those types of data types once your file has been imported and provide content by editing cells directly on your working copy or via a form.

Data Type	Description
Plain Text	UTF-8 encoded text – no formatting.
Formatted Text	UTF-8 encoded text that may contain HTML. Note: HTML will be sanitized to remove any potentially dangerous elements.
Numbers	Numbers should not contain any commas. For negative numbers, the negative sign should precede the number. (e.g. -10000).
Percent	Percent can either be a number or a number followed by a percent sign. However, if it is just a number it should not be in the 0.01 to 1.0 range (e.g. 42 not 0.42).
Money	Money should be a number preceded with a dollar sign. For negative monetary values, either a negative sign or a set of parentheses are acceptable (e.g. $-42.21, ($42.21), or -$42.21).
Date & Time	Supported ISO 8601 formats[14]: • yyyy-MM-dd['T']HH:mm:ssZ (e.g. "1920-01-22T00:00:00Z", "1920-01-22T00:00:00-10:00", or "1920-01-22 00:00:00Z") • yyyy-MM-dd['T']HH:mm:ss (e.g. "1920-01-22T00:00:00" or "1920-01-22 00:00:00") • yyyy-MM-dd['T']HH:mm (e.g. "1920-01-22T00:00") • yyyy-MM-dd (e.g. "1920-01-22") Supported non-ISO 8601 formats: • MMM d, yyyy (e.g. "Feb 19, 1972") • MMMM d, yyyy (e.g. "February 19, 1972") • M-d-yyyy (e.g. "2-19-1972") • M/d/yyyy (e.g. "2/19/1972") • M.d.yyyy (e.g. "2.19.1972") In the non-ISO 8601 formats, months and days can be either single or double digit and may or may not be led with a '0'.
URLs	URL's support two different input formats: • State of Iowa • http://www.iowa.gov/ Only three URL schemes are acceptable: ftp, http, and https.
Email	Three different input formats are acceptable for emails: • State Data Administrator • transparency@iowa.gov • State Data Administrator <transparency@iowa.gov>

[14] The International Standard for the representation of dates and times is ISO 8601. Its full reference number is ISO 8601 : 1988 (E), and its title is "Data elements and interchange formats - Information interchange - Representation of dates and times"

Data Type	Description	
Location	Composite data type based on U.S formatted address information provided in single or multiple columns. Multiple columns: • Street Address • City • State • Zip Code (either 5 or 9 digit) U.S. formatted address in a single column should be comma separated and wrapped in quotes (for CSV) (e.g. "1007 East Grand Avenue, Des Moines, IA, 50319"). The following partial addresses may be geocoded: • City, State • Zip Code • State • State, Zip Code	Location data may also be based on geographic coordinates provided in decimal degrees. Separate columns for lat and long: • Latitude bounded by 90 and -90 • Longitude bounded by 180 and -180 Single column for geographic coordinates should enclose in parentheses, use comma to separate latitude and longitude (latitude listed first) and wrap in quotes (for CSV) (e.g. "(41.591186, -93.603782)"). Addresses and geographic coordinates may also be provided in a single column where wrapped in quotes (e.g. "1007 East Grand Avenue, Des Moines, IA, 50319 (41.591186, -93.603782)")
Checkboxes	Valid false values: • 0 • f • false • n • no • off	Valid true values: • 1 • t • true • y • yes • on
Phone	Is only available after importing a dataset. Number. The system does not validate phone number or format.	
Multiple Choice	Is only available after importing a dataset. System allows dataset owner to pre-enter values to select from a drop-down list.	
Photo	Is only available after importing a dataset. Accepts the following file formats: .jpg, .png, .gif	
Document	Is only available after importing a dataset. Accepts any file type.	

Appendix D. Panel Summary

The following table highlights the functions available in the panels for datasets and views.

Panel	Summary
Edit (datasets only)	The edit panel allows owners of datasets to create working copies of the datasets to edit. On working copies, owners of datasets can append or replace data or add new columns.
Manage	The manage panel allows owners of datasets and views to: 1. Transfer ownership of a dataset or view to another user 2. Delete the dataset or view – ***please note: deleting any dataset will delete all accompanying views*** 3. Sharing the dataset or view with other users. This also facilitates specifying another user as an "owner" 4. Set permissions for the dataset or view – whether it is public or private, and if comments are allowed. Please note that all comments are moderated. 5. Show and hide columns on a dataset – uncheck to hide a column and click "Apply" 6. Change the order of columns within the dataset – drag and drop order – moving a column up will display it further to the left on the display. Once you have ordered the columns, click "Apply"
More Views	The more views panel provides owners access to dataset snapshots, which are previous published versions of your dataset. It also provides a list of views that are based on the dataset.
Filter	The filter panel allows users to: 1. Establishing conditional formatting criteria that can change the background color of rows. 2. Group rows together and summarize data; and sort data based on one or more columns. 3. Filter a dataset based on its contents.
Visualize	The visualize panel provides steps for creating calendars, charts and maps. On existing views – the visualize panel displays options selected to create the view. Users can change these in order to create and save new views.
Export	The export panel provides programmatic access via API, refreshable data link (for datasets only) and multiple static downloads (e.g. CSV, CSV for Excel, JSON, PDF, RDF, RSS, XLS, XLSX, XML). Geospatial data imported using a shapefile can only be downloaded in static formats with geospatial data (e.g. KML, KMZ, or Shapefile) or without geospatial data (e.g. CSV, JSON). Geospatial data provided by a web service has the same export capabilities as other views.
Discuss	The discuss panel allows users to submit a comment related to the dataset or view. All comments are moderated.
Embed	On datasets, the embed panel provides owners the ability to create web forms based on their datasets. It also provides users with html code to embed datasets or views on other websites.
About	The about panel provides information about the dataset or visualization. On this panel, owners of datasets or views are given the ability to edit dataset or visualization's metadata.